www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

EVALUATING A DATA CLUSTERING APPROACH FOR LIFE-CYCLE FACILITY CONTROL

SUBMITTED: June 11, 2012
REVISED: February 20, 2013
PUBLISHED: April 2013 http://www.itcon.org/2013/6
EDITOR: Robert Amor

A. Christopher Bogen, Computer Scientist,
Engineering Research and Development Center, US Army Corps of Engineers;
Chris.Bogen@usace.army.mil

Mahbubur Rashid, Computer Scientist,
Engineering Research and Development Center, US Army Corps of Engineers;
Mahbubur.Rashid@usace.army.mil

E. William East, Civil Engineer
Engineering Research and Development Center, US Army Corps of Engineers;
Bill.W.East@usace.army.mil

James Ross, Computer Scientist
Engineering Research and Development Center, US Army Corps of Engineers;
James.E.Ross@usace.army.mil

SUMMARY: *Data reported by sensors in building automation and control systems is critical for evaluating the as-operated performance of a facility. Typically these systems are designed to support specific control domains, but facility performance analysis requires the fusion of data across these domains. Since a facility may have several disparate, closed-loop systems, resolution of data interoperability issues is a prerequisite to cross-domain data fusion. In previous publications, the authors have proposed an experimental platform for building information fusion where the sensors are reconciled to building information model elements and ultimately to an expected resource utilization schedule. The motivation for this integration is to provide a framework for comparing the as-operated facility with its intended usage patterns. While the authors' data integration framework provides representational tools for integrating BIM and raw sensor data, appropriate computational approaches for normalization, filtering, and pattern extraction methods must be developed to provide a mathematical basis for anomaly detection and "plan" versus "actual" comparisons of resource use. This article presents a computational workflow for categorizing daily resource usage according to a resolution typical of human-specified schedules. Simulated datasets and real datasets are used as the basis for experimental analysis of the authors' approach, and results indicate that the algorithm can produce 90% matching accuracy with noise/variations up to 55%.*

KEYWORDS: *Building Information Modelling (BIM), machine learning, pattern detection, signal processing, building automation and control*

REFERENCE: *A. Christopher Bogen, Mahbubur Rashid, E. William East, James Ross (2013) Evaluating a data clustering approach for life-cycle facility control Journal of Information Technology in Construction (ITcon), Vol. 18, pg. 99 - 118, http://www.itcon.org/2013/6*

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evaluating Data Clustering Approach for Life-Cycle Facility Control	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Corps of Engineers,Engineering Research and Development Center,3909 Halls Ferry Road,Vicksburg,MS,39180-6199**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**20**	

1. INTRODUCTION

Given the current emphasis on sustainability, there is a growing interest in monitoring various facets of building resource consumption. Building monitoring and automation systems most commonly exist as closed-loop systems for security, fire safety, water, electrical, and HVAC (Heating, Ventilation, and Air-Conditioning). Without adequate planning or specific computational goals (e.g. calibrating an energy model), one is quickly overwhelmed by the volume and complexity of data. Thus, data collected from such systems are most commonly used for specific problems such as fault detection, design optimization, and system calibration. Ideally information from each system provides decision support for control of building life cycle cost through continuous commissioning (Liu et. al, 2002), but this ideal requires substantial data integration aligned with a facility life cycle model.

One of the fundamental requirements of a mature engineering process model is the comparison of planned versus actual results for cost, risk, and quality control. For facility engineering this implies that there must be a structured specification of building space requirements and a mechanism for detecting divergent system and occupant behaviours. Such a mechanism would have broad applicability for commissioning, energy efficiency, sustainability, diagnostics, maintenance, and a variety of other problems.

For example, the U.S. federal government (a billion dollar player in the capital facility industry) is attempting to minimize its facility footprint based on the current utilization of its buildings. Such information may not be objectively reported unless there are quantifiable metrics collected and maintained in the context of a mature engineering process. Similarly, there are various mandates for institutional energy reduction and increased sustainability. Often, these problems are addressed through design checklists and energy prediction models that may or may not accurately forecast how the building will be used.

The authors are conducting research to realize life-cycle building control through a model of data exchanges across the entire facility life-cycle and a computational approach for comparing expected and actual resource utilization. This article reports on progress towards the latter effort. The authors hypothesize that application of telemetry noise filtering and a clustering algorithm can provide accurate results for comparing planned and actual daily resource utilization in the context of human-interpretable schedules.

Experiments were conducted to test this hypothesis on an algorithm using simulated and real data and ultimately to determine the algorithm's expected accuracy, sensitivity to noise, and its general applicability. Results reveal that the authors' approach can produce 90% accuracy for three types of waveform variation (intensity, frequency, and shift) at a cumulative signal to noise ratio of 55%.

Before revealing the details of the adopted approach (Section 2), the experimental plan (Section 3), and the experimental results (Section 4), the proceeding subsections highlight related background knowledge and research efforts in architecture, engineering, and construction (AEC).

1.1 Data Mining and Pattern Recognition

One of the effective ways of filtering voluminous data is to discover recurring patterns in the raw data that may be reported to higher order analysis methods. Patterns in telemetry data may be detected using a variety of data mining approaches. Data mining is a mature and active area of research that typically includes the following phases:

- Data dimension and noise reduction: handles noise and normalizes scale, units, or other heterogeneous facets of the data.
- Data set partitioning: organizes data with similar characteristics
- Data classification and labelling: assigns incoming data to categories identified from historical data
- Anomaly detection: applies statistical inference and reasoning to identify novel occurrences of data points

Computational approaches for each phase are chosen according to the constraints of the problem. FIG 1 illustrates the computational methods available for each phase of data mining where right-most items are more computationally demanding. Highlighted items are most relevant to the authors' approach while non-highlighted

items represent related technologies. While there is extensive research in all aspects of data mining, this article focuses on the topics of noise reduction and clustering (relevant to multiple data mining phases).

Interpolation (e.g., cubic splines) is a noise reduction method for supplying missing data in the data set (Vaseghi, 2009). Essentially such an interpolation technique is used to smooth out jumps in the original data set. The cost of interpolation is *O(n)* where *n* is the size of the input data set.

Discrete Fourier Transformation (DFT) is used to transform a time-domain input signal into its frequency-domain components. Using this transformation we can detect and eliminate unwanted frequency components. Taking the inverse DFT of the modified frequency-domain representation will yield a relatively less noisy data set. A naïve implementation of the DFT has a cost of $O(n^2)$ where *n* is the input data size. Using the *Fast Fourier Transformation* (FFT) algorithm (Cooley & Tukey, 1965) may reduce this cost to *O(nlogn)*. However, in both the naïve and FFT the cost of the algorithm is directly proportional to *n*, the input data size.

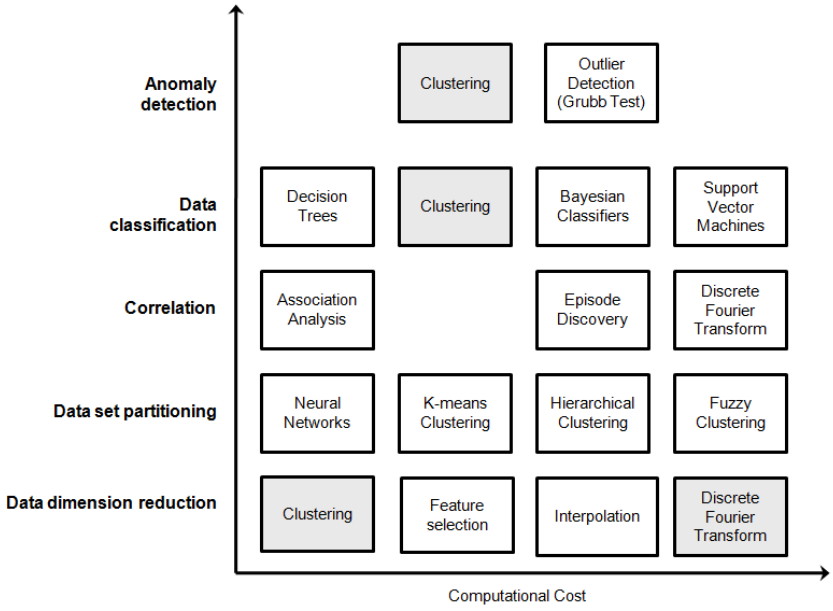

FIG 1: Data Mining Approaches

Clustering is used throughout the data mining process to detect similarities between data points. In clustering, a score function measures similarity between two data points. Common approaches to score functions include a single linkage, complete linkage, and centroid comparison. Each scoring function has variable performance depending on the characteristics of the dataset (Saitta, Raphael, & Smith, 2007).

A popular method for data partitioning is *k-means clustering* (MacQueen, 1967). In this technique the data items are partitioned into *k* number of sets so that the resulting *intra*-cluster similarity is high but the *inter*-cluster similarity is low. The measure of similarity used in this technique is based on the different types of *norms* (such as, *Euclidean*, *infinite*, etc.) between the mean value of a cluster and the value of a particular data item in the cluster. Each of the mean values is considered a central value of that particular cluster. The number of partitions, *k* is selected by the user. The total cost of this technique is *O(nkt)*, where *n* is the number of input data items, *k* is the number of intended partitions, and *t* is the number of iterations required to reach below a specified cluster threshold.

A variant of the clustering algorithm is the *hierarchical clustering* (Day & Edelsbrunner, 1984) where clusters are progressively *merged* (bottom-up) or *split* (top-bottom). The main problem with this type of partitioning technique is

that a poor choice of the decision to merge or split the clusters cannot be remedied by backtracking. This results in poor partitioning of the input data items. To correct any particularly derogatory situation the whole process of merging must restart from the beginning. Another problem faced in hierarchical clustering is the scalability problem. This is because the algorithm must inspect a large number of data points to calculate a particular cluster means. The computation load directly depends on the number of cluster means that need to be calculated. As a result data items with wide variances will cause poor scalability in the algorithm.

Fuzzy clustering is a portioning method (Cook, 2007) whereby a single data point may belong to more than one cluster. Each data point is assigned a set of membership level values. These values indicate the strength of the membership of the data point with each of the clusters.

1.2 Related Work in Architecture, Engineering, and Construction

Sensor data fusion, analysis, and visualization is an active and frequent topic in the Architecture-Engineering-Construction (AEC) research community (Shahandashti et al., 2011) for a variety of applications including: non-intrusive load monitoring (Bergés et al., 2008), mashups of open-source BIM and sensor data (Zach et al., 2012), Energy Analysis (Kim et al., 2011) (Ahmed et al., 2011) (Mail et al., 2012), Indoor localization (Pradhan et al., 2009), tracking of resources on construction sites (Song et al., 2006) (Park et al., 2012), structural health monitoring (Posenato, et al. 2008), deriving as-built geometry models from point cloud data (Tang et al., 2010), and deployment of integrated, campus-wide automation and monitoring systems (Rowe et al,. 2011). Researchers are also quantifying the influences of occupant behaviour on facility resource consumption (Yu et al., 2011).

Works such as those produced by Posenato et al. (2008) and Pradhan et al. (2009) offer experimental evidence for the utility clustering algorithms in AEC applications. Posenato et al. (2008) presentation of an approach to long-term monitoring of a complex structure provides sound evidence for the utility of clustering at a level of abstraction above specific structural models or modes of analysis. Likewise, works by Kim et al. (2011) and Ahmed et al. (2011) demonstrates the utility of data mining approaches for analysis of energy consumption data.

Realizing the full potential of AEC sensor data analysis requires integration into a building life cycle populated by interoperable data sources and building information models. Building performance modelling and simulation efforts such as those by Maile et al. (2012) and Bazjanac (2012) begin to fill research gaps in comparative performance analysis between design and commissioning. Enabling technologies such as the Monitoring System Toolkit (MOST) (Zach et al. 2012) and Sensor Andrew (Rowe et al. 2011) contribute technologies that reconcile disparate sensor technologies and building information models.

2. APPROACH

The authors have developed a life-cycle information exchange for the entire life of a facility (East, Love, & Nisbet 2010). This life cycle information exchange model is based on the Industry Foundation Class (IFC) and the related Construction Operators Building information exchange (COBie) international building information model standards. This life-cycle information exchange model specifies the IFC/COBie data exchanges that may originate from a variety of perspectives such as HVAC, water, electric, production selection, design handover, etc. Two elements particularly relevant to this article are: (1) Building Programming information exchange (BPie), and (2) Building Automation Monitoring information exchange (BAMie) (East 2012 a,b).

Building Programming information exchange provides specifications for expected resource use. Building space planning is an early life cycle activity when the building may be identified in terms of various requirements for capacity, service, schedules, and conformance to relevant policies. In some instances, design guides specify such requirements for a variety of facility and space types (Unified Facilities Criteria Program (UFC), National Institute of Building Sciences (NIBS), 2012).

The Building Automation Monitoring information exchange is a draft model view definition that describes how the IFC model should be used to specify building automation system product information, physical and logical connections to other design elements, data point addressing, performance history, and device configuration. In

short, this model view definition provides a necessary connection between sensor data about a building and various models of the building – geometrical, functional, structural, electrical, managerial, etc.

Given the life cycle information exchange integration between facility requirements and sensor data, it is possible to compare expected and actual utilization of facility resources (East, Bogen, & Rashid, 2012). To this end, the authors developed and evaluated an algorithm that categorizes daily utilization data and may be compared to expected resource utilization schedules at an appropriate resolution. Algorithm design decisions were based on considerations about resource utilization data and a general representation of expected/actual data.

2.1 Resource Utilization Data Considerations

When a facility is used its occupants execute various activities according to schedules that have a relatively low resolution when compared to the possible resolutions available in sensor data. For example, the common occupant schedules for work shifts and meetings are not typically scheduled to occur and end on the 11th minute, 30th second of an hour. Instead events and activities typically occur within quarter hour increments unless such activities require extraneous time precision. Likewise, the target resolution of incoming sensor data for the authors' work is fifteen minutes. While 15-minute data may be suitable for analysing schedules and resource consumption, it is not suitable for more intricate tasks such as fault detection or structural health analysis.

Scheduling of resources (by persons) is also typically done in binary units where resources are active/inactive, available/not available. In contrast, raw sensor data is typically continuous values derived from analog control voltages. Likewise, the measured resource may be subject to variations that are inherent to human behaviour. For example, consider a light level sensor in a conference room that has been scheduled for team meetings between 3:00pm and 5:00pm, Monday through Wednesday. If meetings occur on all days it is unlikely that they will start and finish precisely according to schedule. Also, during a longer meeting, there is more chance of variability in the frequency of resource use over the entire meeting – e.g. if the meeting involves multi-media presentations then lighting may be dimmed, or if it is a long meeting then people are likely to take extended breaks and turn off the lights.

These variations of humans, resource systems, and sensors may be considered as noise when compared to expected resource schedules. When designing the algorithm the authors considered the following categories of noise (illustrated in FIG 2):

- *Intensity Noise*: This type of noise occurs when the sensor output varies from its nominal expected value.
- *Shift Noise*: This type of noise occurs when an expected event occurs earlier or later than expected.
- *Frequency Noise:* This type of noise occurs when the occurrence of utilization indicators fluctuate over the duration of expected resource use.

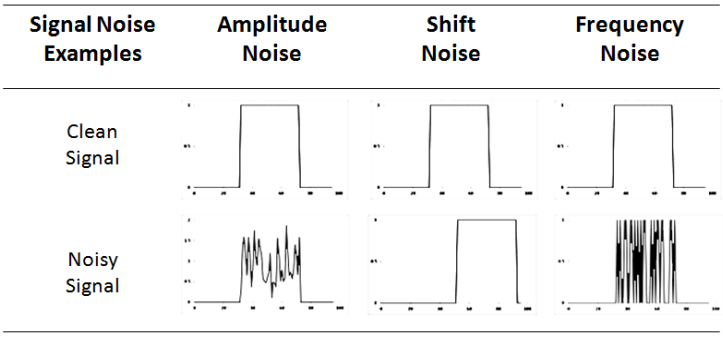

FIG 2: Examples of Intensity, Shift, and Frequency Noise

2.2 Reference Patterns

The authors created an initial set of reference patterns that were used during development testing and in later experiments (Section 3). These patterns specify the binary consumption pattern of a resource for an *8 hour workday*, *9 hour workday*, *10 hour workday*, *12 hour work day*, *15 hour workday*, *18 hour workday*, *21 hour workday*, *Constant use*, *Constant non-use*, *3 peak workday*, *2 peak workday*. FIG 3 illustrates example reference patterns for an 8 hour workday, a 2 peak workday with a lunch break, a 3 peak workday, and a 12 hour workday. When compared with actual resource waveforms these patterns represent the usage characteristics of long duration use electrical devices such as lighting system, desktop computers, or exhaust fans. Typically, these devices stay operational (on) for long intervals during facility occupancy and remain inoperative (off) otherwise.

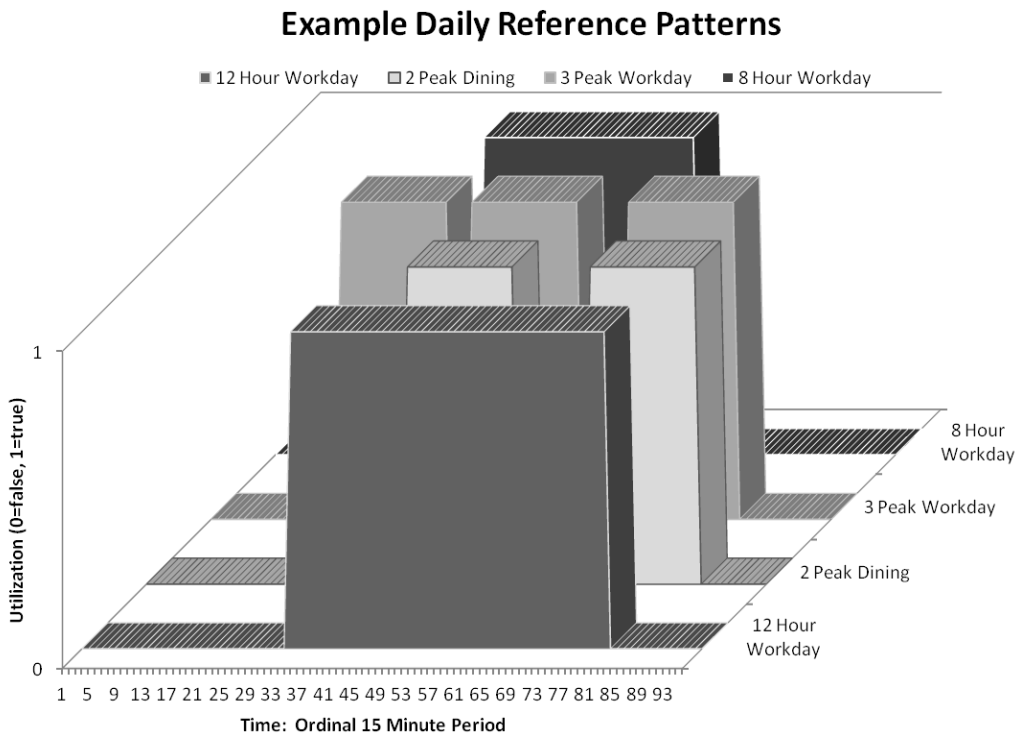

FIG 3: Example Daily Reference Patterns

2.3 Noise Reduction and Pattern Detection Algorithms

The authors have developed an approach that implements the first two phases of data mining (data dimension/noise reduction and data set partitioning) to extract patterns in observed telemetry comparable to the reference patterns (Section 2.2). FIG 4 illustrates the organization of the different components of the processing workflow where various noise and variations are filtered and data is categorized into patterns. The blocks in the upper portion of FIG 4 operate on intensity noise while the blocks in the lower portion of FIG 4 operate on frequency and intensity noise. The following subsections describe these approaches in more detail.

2.3.1 Intensity Noise Reduction

The intensity noise reduction algorithm consists of three steps: (1) Fast Fourier Transformation (FFT), (2) Spectral Subtraction, and (3) Inverse Fast Fourier Transformation. The Fast Fourier Transformation algorithm closely follows the one described by Cooley and Tukey (1965) and implemented by Press et al.(2007). The Fast Fourier Transformation is a popular algorithm that implements discrete Fourier transformation.

Spectral Subtraction is a technique that has been used primarily for noise reduction in audio signals. One of the issues in the application of Spectral Subtraction for audio is that even though the signal's original noise contents are reduced, the technique causes the introduction of a different type of noise in the data. This noise is manifested as rumbling in the background of the audio signal. This disturbance may become a problem in audio signal processing because it may interfere with audio communication. However, the rumbling is of less concern in resource usage signals since in this case we are concerned only with detecting the transient pulses in the signal.

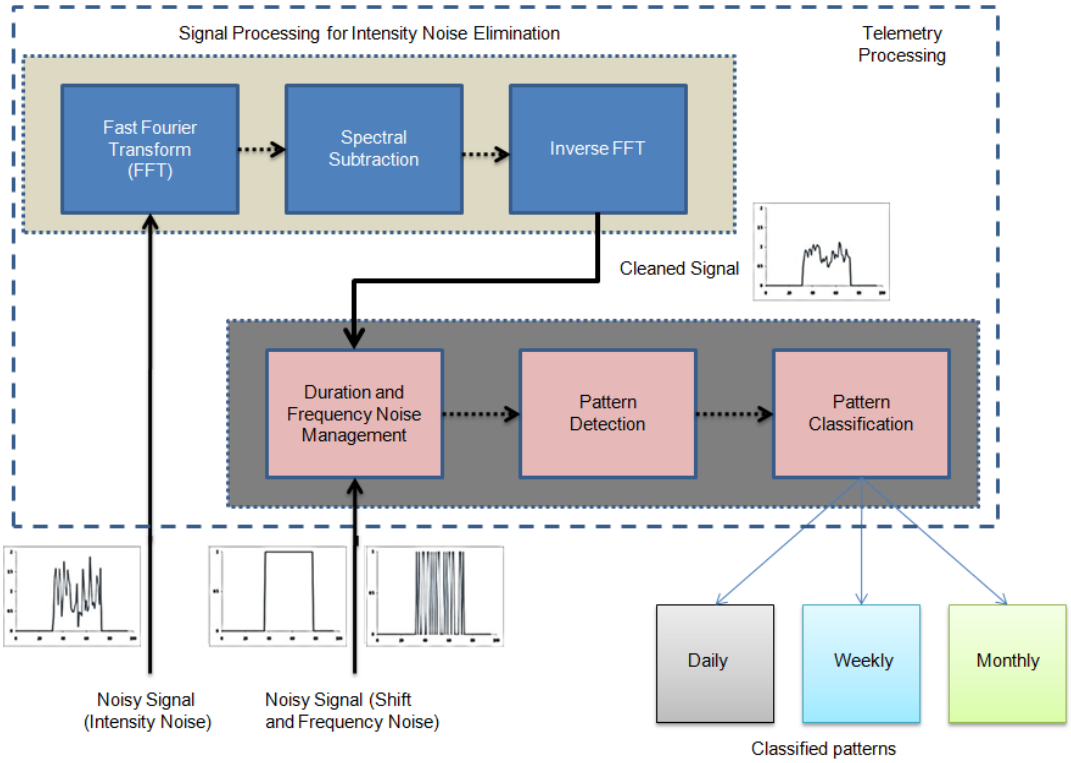

FIG 4: Architectural view of the authors' noise reduction and pattern classification approach

2.3.2 Pattern Extraction and Dataset Partitioning

Frequency and shift variations are addressed by application of an unsupervised *k-means* clustering algorithm. In this approach, each clustered data point is a one day waveform of 15-minute resolution data. Cluster analysis can recognize patterns without any a priori knowledge about the distribution or organization of data. Essentially, the outliers are clustered into separate classes based on their dissimilarities from one another. The clustering algorithm is augmented by an initial population of expected patterns for common office building daily work schedule. Significant deviations from these patterns will result in the creation of a new cluster. However, the clustering algorithm can also create categories without an initial population of patterns.

The function of the algorithm is controlled via a *classification* threshold. This threshold establishes the maximum distance between two data points in the same cluster. The distance between the clusters is calculated by the *root mean squared* (*RMS)* or the *geometric distance* between the daily patterns. RMS is a common, relatively simple classification threshold that serves as a baseline approach for this study. The distance is calculated according to Equation 1, where, a_i and b_i are individual data points of the daily patterns a and b, respectively. N is the number of samples in a data point. In this case, N is 96 (24 hours * 4 sensor readings per hour).

If the minimum RMS distance between a data point and any existing cluster is above the threshold then an average distance measure is calculated and used for a "last chance" match. If the minimum average distance measure between the data point and existing pattern is greater than the threshold then a new cluster is created. Equation 2 provides a specification of the average distance measure where, a_i and b_i are individual data points of the daily patterns *a* and *b*, respectively and N is the total number of samples in a data point.

$$rms = \sqrt{\frac{1}{N}\sum_{N}(a_i - b_i)^2}$$

Equation 1

$$avg = \frac{1}{N}\sum_{i}\left|(a_i - b_i)\right|$$

Equation 2

3. EXPERIMENTAL SETUP

The authors sought to determine the accuracy of the computational approach, its ability to deal with noise, the influence of model parameters, and ultimately its performance on real resource utilization data from facility sensors. These goals were partitioned into two experimental stages: (1) Identifying the solution space with simulated data, and (2) Comparing cluster performance on real facility utilization data.

3.1 Identifying the Solution Space with Simulated Data

The initial goal of the experiments is to approximate the "breaking points" of the algorithm by observing its performance under controlled conditions. First, the daily reference patterns were organized into weekly schedules, and weekly schedules were distributed among the months of the year. Expected weekly and daily pattern sequences, stored in XML documents, were used as criteria for determining accuracy – if the algorithm matches a data point occurrence to the expected data point classification then the occurrence is expected and the algorithm produced an accurate result, otherwise the result is unexpected and the algorithm produced an incorrect result.

An experimental data generator was implemented to introduce specified noise amounts into the reference patterns. The noise generators were executed sequentially as follows:

1. Frequency Noise: the utilization values (0 or 1) are flipped, and the number of points flipped is equal to the ceiling value of the quantity [(Noise %) * (the number of data points in a day)]. For example, a 90% frequency noise amount means that 87 of the 96 binary data points will be flipped, and the index of these points are randomly selected.
2. Intensity Noise: the signal to noise ratio specifies a standard deviation that is applied to all raw data points. For example, for 30% intensity noise, each raw data point is multiplied by a random number between 0 and 0.30.
3. Shift Noise: the start of the peak period in the data is offset proportional to the duration of the activity period. For example, 25% shift noise of an 8 hour workday starting at 6am means that the start or end time may be shifted by up to 2 hours/ (8 15-minute periods).

Performance results were collected for all combinations of intensity, frequency, and shift noise for 5% increments – 9,261 combinations total. For each combination of noise parameters the cumulative accuracy was calculated over a

2-year period for cluster thresholds 0-100% in 5% increments. For debugging purposes, the accuracy of each individual daily pattern was recorded.

3.2 Comparison of Simulated and Real Data

The goal of this stage of the experiments is to determine the applicable accuracy of the algorithm and determine realistic noise levels. Since an expected resource utilization schedule is not available for this building, performance results were approximated by aggregating experimental results from the first stage of the experiment (Section 3.1) with results collected from real sensor data.

Researchers at Penn State University shared approximately eight months of data collected in a commercial office building instrumented for an extensive Department of Energy funded research grant. This data serves as the real data that was compared to the simulated results discussed in Section 3.1.

The following five data points were selected from the dataset: (1) whole building electrical consumption, (2) lighting systems electrical consumption, (3) main office lighting level sensor, (4) air-handling unit electrical consumption, and (5) cooling-unit electrical consumption.

These data points were consumed by the noise reduction/pattern matching algorithm and given an initial set of daily reference patterns. Clustering results were recorded for cluster thresholds from 1 to 100% in 1% increments. The goal of this activity was to determine the minimum cluster threshold that produces no more than one generation of variant clusters.

This minimum threshold identifies a compromise between precision and complexity – i.e. the algorithm should discriminate between reference patterns, but not overwhelm the user with hundreds of slightly variant anomalies. It is assumed that real buildings and their resources will have their own unique variant of our reference patterns, and the algorithm should provide sufficient allowances for the discovery of those patterns. Then, any deviations beyond the first generation of variants may be considered anomalous.

After collecting clustering results from the real sensor data, results were collected from the artificial data using the minimum thresholds discovered from the real data. Then the simulated data results were filtered to eliminate noise levels that result in more than one variant for any reference pattern and produce an accuracy of 90% or greater. The noise levels and accuracy results of the remaining data points represents the expected noise and accuracy of the algorithm on data points comparable to those in the real sensor data points.

4. EXPERIMENTAL RESULTS

Results from the experiments reveal that accuracies (matches to reference or 1st generation variants) of 90% and greater may be achieved through various combinations of noise where frequency noise is between 0-30%, intensity noise is 0-40%, and shift noise is 0-5%. More detailed analysis of each experimental stage is presented in the following sub-sections.

4.1 The Solution Space (Simulated Data)

The data obtained in the experiments was analysed to estimate the accuracy of the detection capability of the pattern detection algorithm. The accuracy is calculated using Equation 3 where *d'* is the number of expected patterns successfully matched in the test data, and *d* is the number of total data points (days) in the data set. This strict measure of accuracy is used to determine the points when noise distorts a reference pattern enough to create a new cluster or match to another cluster.

$$accuracy = \frac{d'}{d}$$

Equation 3

First, the entire solution space is illustrated by creating a 3D surface chart (FIG 5) where the x-axis (horizontal) represents noise combinations ranked by accuracy from largest to smallest. The y-axis represents the cluster threshold and the z-axis (vertical) represents accuracy. The accuracy peak between 90-100% is a light shaded area in the top left of the chart between the x-axis values 0-715.

Surface charts are also provided for each isolated noise element from 0 to 100% in 5% steps. Intensity noise has the least influence on accuracy while shift noise has the highest influence on accuracy. FIG 6 illustrates accuracy for intensity noise, FIG 7 illustrates accuracy for only frequency noise, and FIG 8 illustrates accuracy for only shift noise.

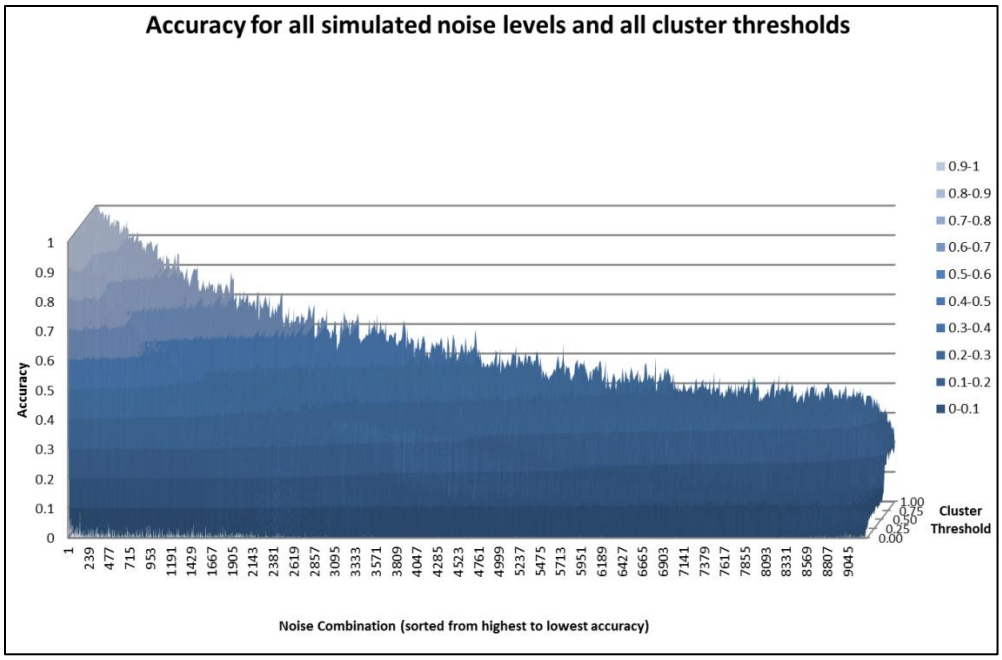

FIG 5: Accuracy for all simulated noise levels and all cluster thresholds

Given a minimum to moderate amount of noise the accuracy typically increases as the cluster threshold increases. However, this is not true for shift noise. The algorithm by default checks distances between reference patterns and the test pattern using the RMS distance measure. This distance measure is especially discriminating for shift noise, but average distance measure is less discriminating for shift and frequency noise. This is due to the fact that average distance measure tracks the shape of the wave forms and is able to detect similar shaped wave-forms even in the presence of time shifts (shift noise) and missing data points (frequency noise). Table 1 and Table 2 provide a cross reference of distances between the reference patterns for RMS and average distances respectively. Note that average distances are typically less than RMS distances. If the cluster threshold is large enough then the RMS distances will provide a match in which case accuracy begins to decline for shift noise.

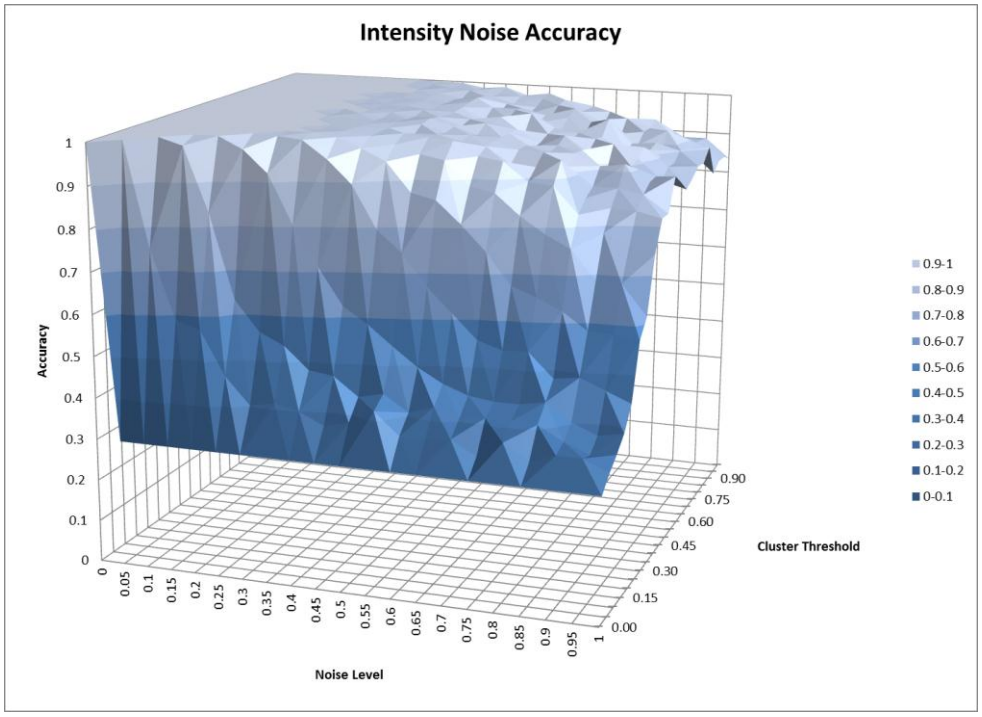

FIG 6: Accuracy in presence of only intensity noise

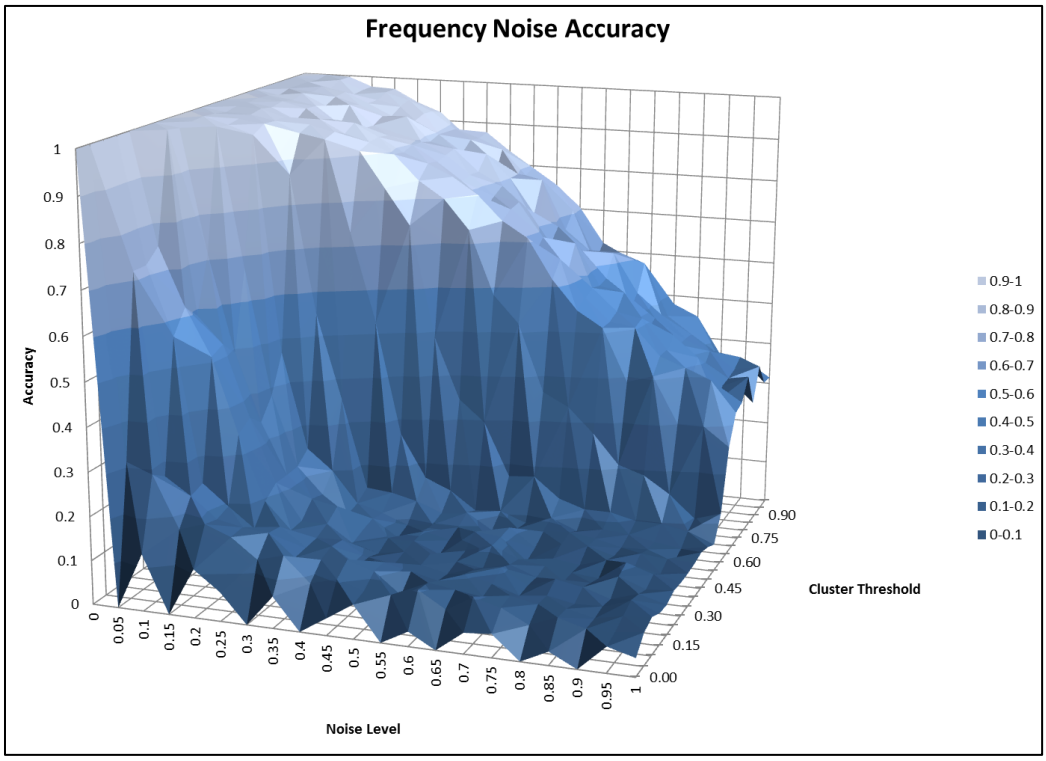

FIG 7: Accuracy in the presence of only frequency noise

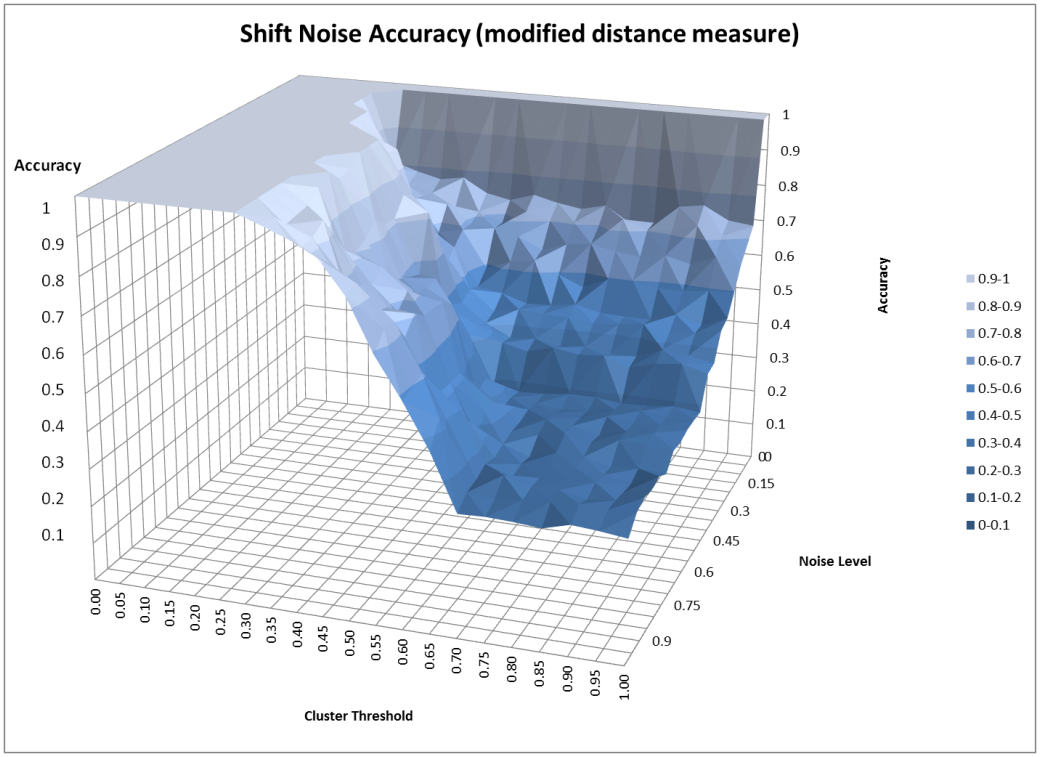

FIG 8: Accuracy in the presence of only shift noise

Table 1: RMS distances between reference patterns (less than 0.10 highlighted)

	10Hr	12Hr	15Hr	18Hr	21Hr	2PD	2PND	3P	8Hr	9Hr	ConstH	ConstL
10Hr	0.00	0.29	0.46	0.58	0.68	0.57	0.64	0.56	0.29	0.20	0.76	0.65
12Hr	0.29	0.00	0.35	0.50	0.61	0.49	0.57	0.48	0.41	0.35	0.70	0.71
15Hr	0.46	0.35	0.00	0.35	0.50	0.53	0.44	0.43	0.54	0.50	0.60	0.80
18Hr	0.58	0.50	0.35	0.00	0.35	0.64	0.57	0.48	0.65	0.61	0.49	0.87
21Hr	0.68	0.61	0.50	0.35	0.00	0.73	0.67	0.60	0.74	0.71	0.34	0.94
2PD	0.57	0.49	0.53	0.64	0.73	0.00	0.58	0.51	0.59	0.60	0.80	0.60
2PND	0.64	0.57	0.44	0.57	0.67	0.58	0.00	0.57	0.70	0.67	0.75	0.66
3P	0.56	0.48	0.43	0.48	0.60	0.51	0.57	0.00	0.63	0.60	0.68	0.73
8Hr	0.29	0.41	0.54	0.65	0.74	0.59	0.70	0.63	0.00	0.20	0.81	0.59
9Hr	0.20	0.35	0.50	0.61	0.71	0.60	0.67	0.60	0.20	0.00	0.78	0.62
ConstH	0.76	0.70	0.60	0.49	0.34	0.80	0.75	0.68	0.81	0.78	0.00	1.00
ConstL	0.65	0.71	0.80	0.87	0.94	0.60	0.66	0.73	0.59	0.62	1.00	0.00

Table 2: Average distances between reference patterns (less than 0.10 highlighted)

	10Hr	12Hr	15Hr	18Hr	21Hr	2PD	2PND	3P	8Hr	9Hr	ConstH	ConstL
10Hr	0.00	0.08	0.21	0.33	0.46	0.07	0.01	0.10	0.08	0.04	0.57	0.43
12Hr	0.08	0.00	0.13	0.25	0.38	0.16	0.07	0.02	0.17	0.13	0.49	0.51
15Hr	0.21	0.13	0.00	0.13	0.25	0.28	0.20	0.10	0.29	0.25	0.36	0.64
18Hr	0.33	0.25	0.13	0.00	0.13	0.41	0.32	0.23	0.42	0.38	0.24	0.76
21Hr	0.46	0.38	0.25	0.13	0.00	0.53	0.45	0.35	0.54	0.50	0.11	0.89
2PD	0.07	0.16	0.28	0.41	0.53	0.00	0.08	0.18	0.01	0.03	0.65	0.35
2PND	0.01	0.07	0.20	0.32	0.45	0.08	0.00	0.09	0.09	0.05	0.56	0.44
3P	0.10	0.02	0.10	0.23	0.35	0.18	0.09	0.00	0.19	0.15	0.47	0.53
8Hr	0.08	0.17	0.29	0.42	0.54	0.01	0.09	0.19	0.00	0.04	0.66	0.34
9Hr	0.04	0.13	0.25	0.38	0.50	0.03	0.05	0.15	0.04	0.00	0.61	0.39
ConstH	0.57	0.49	0.36	0.24	0.11	0.65	0.56	0.47	0.66	0.61	0.00	1.00
ConstL	0.43	0.51	0.64	0.76	0.89	0.35	0.44	0.53	0.34	0.39	1.00	0.00

By examining the plots for individual noise elements a large portion of the solution space peak was scooped by filtering for shift noise less than or equal to 10%, frequency noise less than or equal to 55%, and intensity noise less than or equal to 100%. FIG 9 illustrates the resulting accuracies for the remaining (after filtering) noise combinations and cluster thresholds.

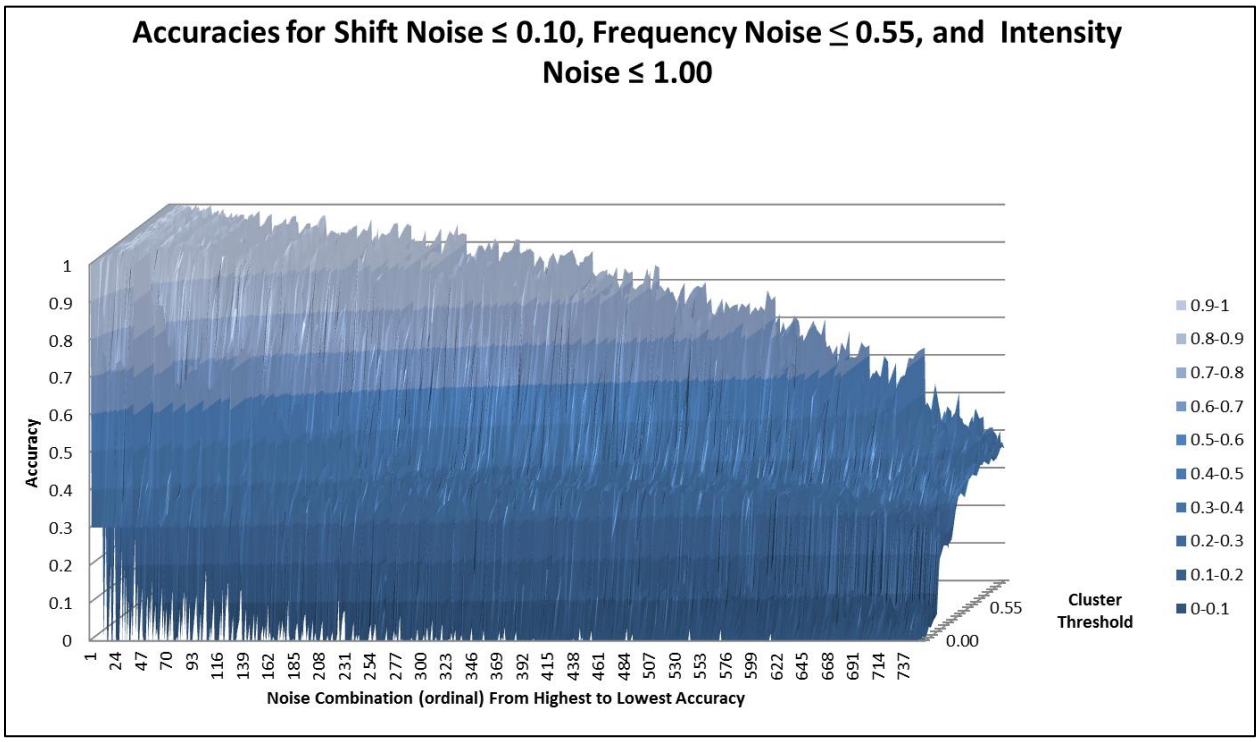

FIG 9: Accuracy for shift noise ≤ 0.10, frequency noise ≤ 0.55, and intensity noise ≤ 1.00

4.2 Real Data Results

Various facets of the real data were examined to ensure quality of the underlying data, to observe clustering performance, and collect data for comparative analysis. The minimum cluster thresholds that produced no more than one cluster variant for any reference pattern are between 6 and 9% for the 5 data points. Table 3 summarizes the clustering performance of the algorithm on the 5 real data points for the minimum cluster threshold that produces

no more than one generation of cluster variants. For each data point, the following information is listed: the cluster threshold, total number of variant clusters derived, distinct count of observed daily patterns, the number of data points, and the ratio of variant clusters to total data points.

It was unexpected that the main office lighting level sensor ranked with the lowest variation as its readings are subject to the variation of a relatively small population when compared to lighting for the entire building. One possible explanation is that the pattern of ambient light from windows may provide some regularity to the data – assuming that it is not an overcast day. It was, however, expected that the cooling and air handling units would have substantial variation due to seasonal temperatures changing between February and September, and this is reflected in the reported statistics.

Table 3: Cluster Counts for Real Data Sorted by % of Variant Cluster Occurrences

DataPoint	Cluster Threshold	Total Variant Clusters	Distinct Cluster Occurrences	Day Count	% Variant Clusters
Office Light Level	9%	1	7	170	17.06%
All Lighting (electric)	6%	3	9	211	20.85%
Air Handling Unit (electric)	6%	2	10	211	27.96%
Whole Building (electric)	7%	2	8	211	38.86%
Cooling Unit (electric)	7%	2	11	211	60.19%

Examining the plots of various daily sequences provides additional insight about the variation of data and the algorithm's performance while also providing indicators for the quality of the data. FIG 10 illustrates main office lighting data and the matched reference pattern from April 20-24, 2012. This sequence of five of the seven distinct clusters occurring in these data points. Note the difference in time between the start of the 12 hour reference pattern peak and the start of its corresponding raw data. This is an example of the algorithm allowing for a margin of variability of shift noise, and likewise, the variation in amplitudes is an example of intensity noise. The data in FIG 10 seems to indicate that the office was not occupied on April 23 and 24.

The all lighting (electric) data illustrated in FIG 11 also indicate there was decreased lighting consumption on April 23 and 24. Essential lighting for hallways, stairwells, lobbies, and exterior likely accounts for the two 8-hour variant patterns matched on those days. If this technology was fielded it would be necessary for practitioners to annotate such variants if they are common recurrences – e.g. 32 of the 211 days of all lighting data points are clustered to the 8 hour variant.

Inspection of seasonal performance of the whole building and the chiller unit reveals an expected correlation between summer months and chiller unit consumption, and ultimately, chiller unit consumption and total electric consumption. FIG 12 and FIG 13 illustrate this correlation for the whole building and chiller unit energy consumption on days from February, March, and July. Note that the correlation between chiller unit and total electric consumption is strongest during the summer months.

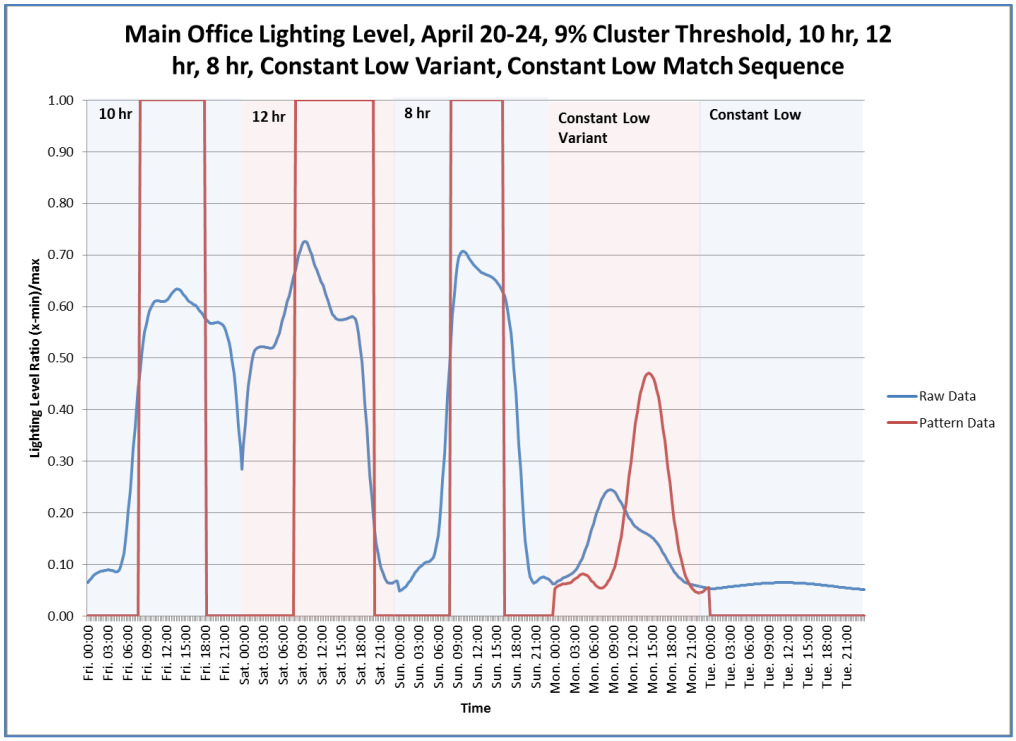

FIG 10: Sequence of pattern matches for main office lighting level

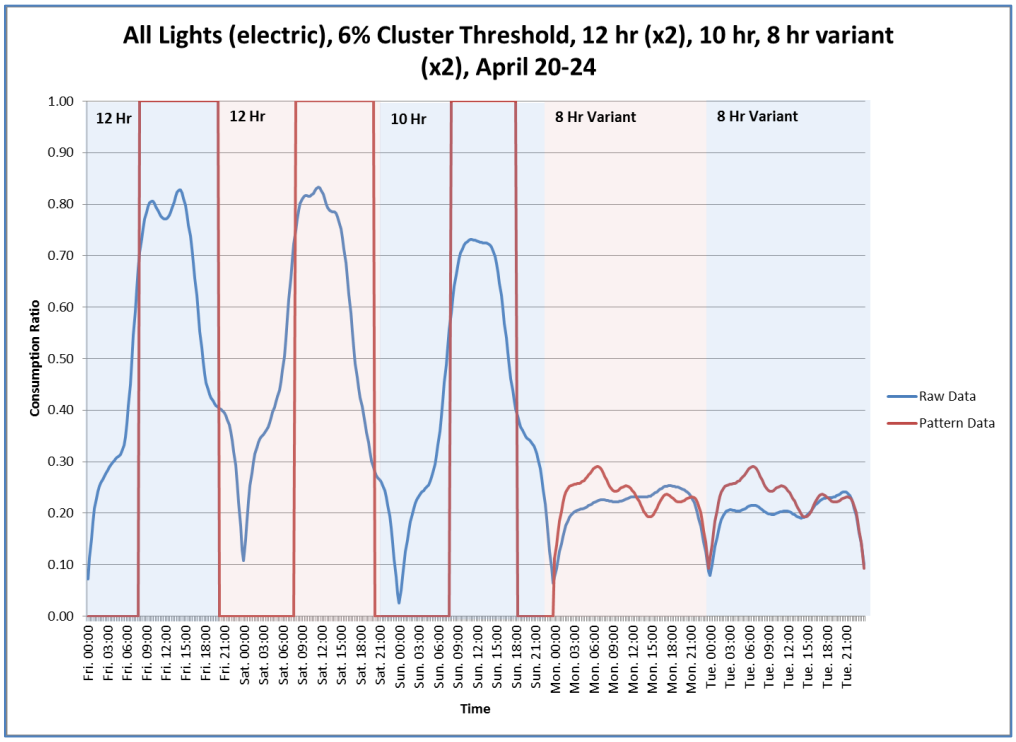

FIG 11: Sequence of pattern matches for all lighting (electric)

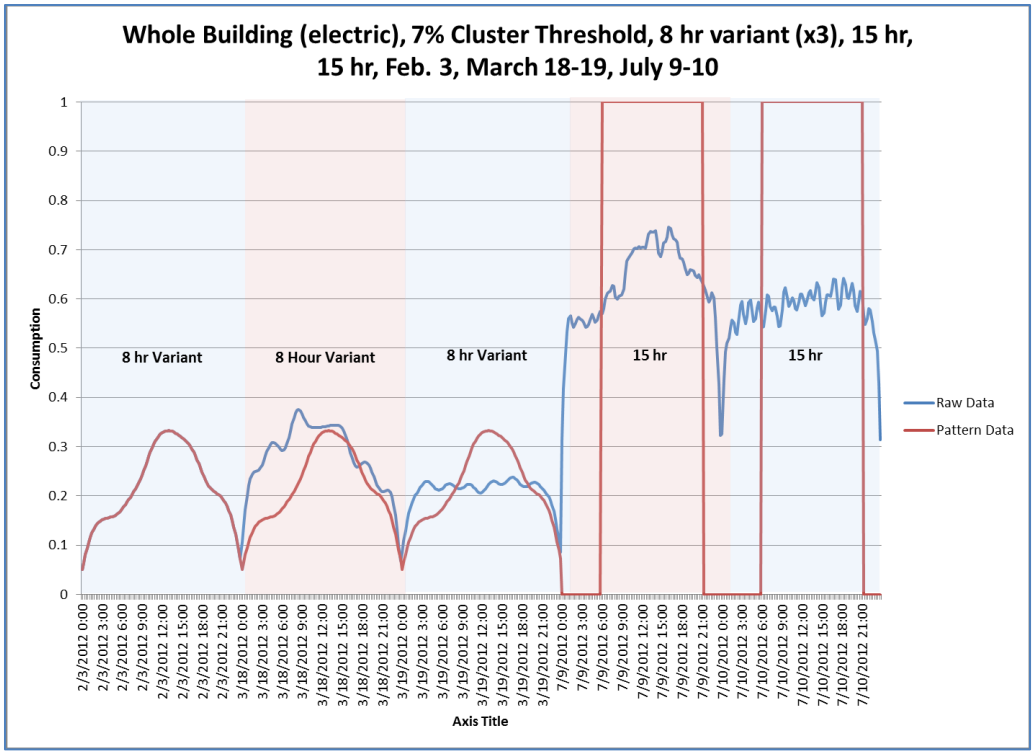

FIG 12: Whole building (electric) matches from Feb., March, July

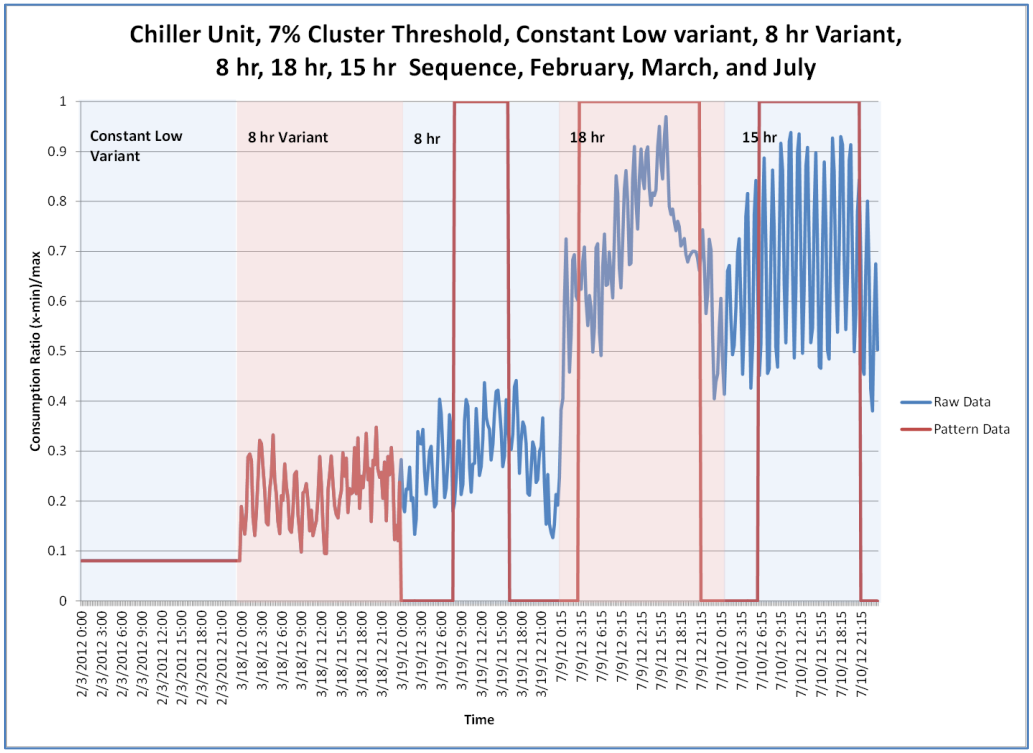

FIG 13: Chiller unit example pattern matches from Feb., March, July

4.3 Comparing Real and Simulated Data

After deriving the minimum cluster thresholds for the real data points (see Table 3), additional data was collected using simulated data. The reference pattern population was trimmed to exclude patterns not occurring in the real data - constant high, 3 peak, and 2 peak days. The criteria for an expected match were extended to include matches on first generation variants.

Once accuracy values were obtained for all noise settings, the noise settings producing more than 1 variant for any reference pattern or producing accuracy less than 90% were eliminated. Examining the data left for the lower and upper bounds of the real data cluster thresholds (6% and 9% respectively) provide an approximation of the noise level boundaries that may produce accurate results on the real data points – and for similar facility data points.

FIG 14 illustrates the results for the lower bound, 6% cluster threshold, and FIG 15 illustrates the results for the upper bound, 9% cluster threshold. While various combinations of noise produce accurate results, the range for individual noise components is 0-40% intensity noise, 0-30% frequency noise, and 0-5% shift noise. The maximum sum of total noise is 55%.

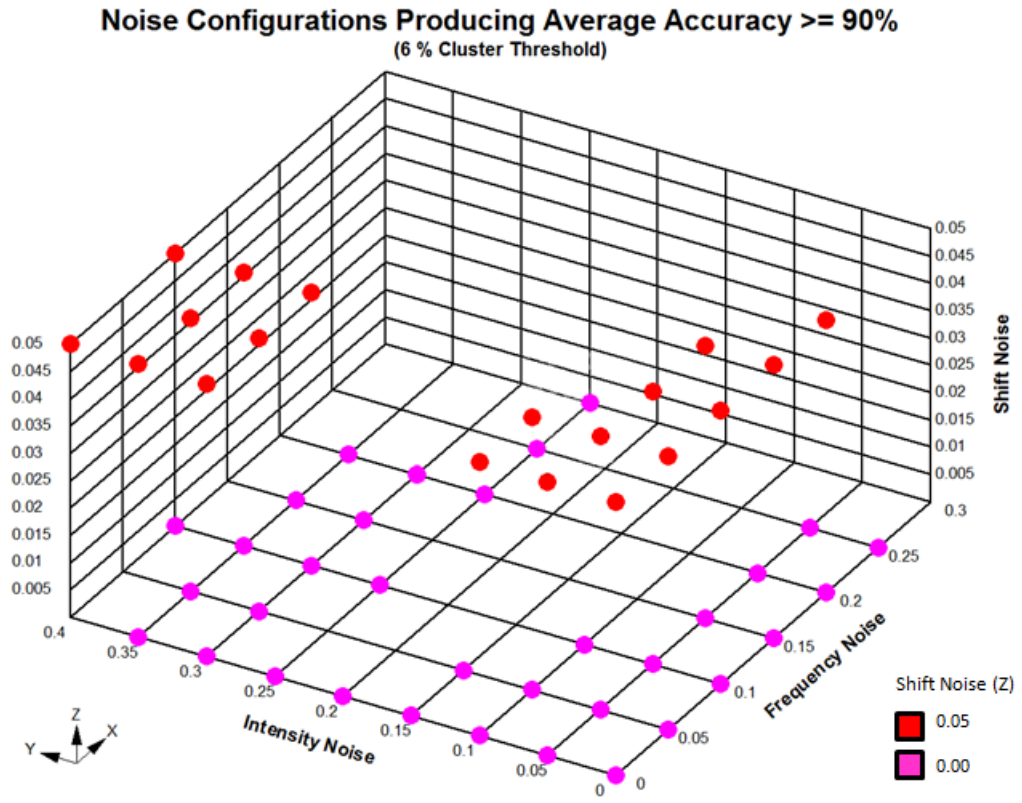

FIG 14: Noise configurations producing accuracy >= 90% for 6% cluster threshold (minimum for real data)

5. CONCLUSIONS AND DISCUSSION

K-means clustering is sufficiently accurate for comparing expected and actual resource consumption schedules when combined with a series of pre-processing noise reduction algorithms: Fast Fourier Transform, Spectral Subtraction, and Inverse Fast Fourier Transform. The approximation of accuracy between artificial and real data provides a prediction of noise levels in the real data as well as an approximation of expected accuracy. While the algorithm was

evaluated for a multitude of artificial noise conditions, the artificial noise seems unrealistically disruptive when compared to the relatively clean recurring patterns observed in the real sensor data.

The artificial data experiments revealed quite a lot of poor performance on various configurations of noise, but this does not necessarily indicate poor overall performance. The noise configurations were enumerated as a necessary measure to determine the amount of disruption that was necessary to create variant clusters and incorrect matches. Since the injection of noise is random, it essentially creates new patterns that could be closer to another reference pattern than its original noise-free ancestor. Such possibilities are compounded by a diverse population of reference patterns that included pairs that were far and close to one another – thus raising the potential for incorrect mappings.

While the experiments and approach focused on evaluation of noise-reduction and clustering, the adopted approach provides rudimentary classification and anomaly detection capabilities – e.g. if more than two consecutive weeks exhibit unexpected behaviour then a smart resource utilization system may prompt a facility manager to acknowledge or deny an alert. When thinking of energy consumption it is common to think of alarms being raised when consumption is over a specific threshold. Flagging anomalies based on expected schedules of use means that alarms may also be raised when rooms and resources are underutilized.

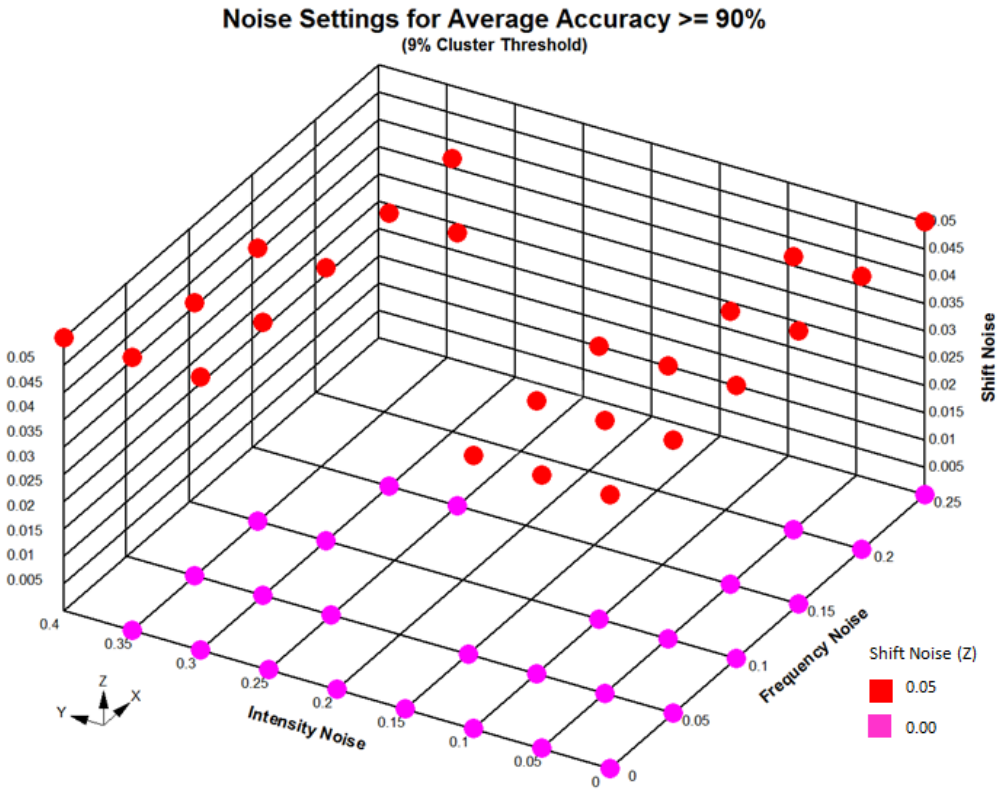

FIG 15: Noise configurations producing accuracy >= 90% for 9% cluster threshold (maximum for real data)

The proposed approach is beneficial because it requires low resolution, unit-neutral data that is not likely to place additional constraints on the sampling programs of installed building automation and monitoring systems – reprogramming of data logging equipment may be cost prohibitive in some circumstances. However, further research must be performed to demonstrate the authors' full facility life-cycle control concept. These computational methods must be evaluated in the context of a real-facility described by the authors' life-cycle information exchange model.

6. ACKNOWLEDGEMENTS

The authors wish to thank John Messner and Scott Wagner, Penn State University, for their generous sharing of sensor data that was used in this article.

7. REFERENCES

Ahmed A., Korres N., Ploennigs J., Elhadi H., and Menzel K., Mining Building Performance Data for Energy-Efficient Operation, Advanced Engineering Informatics, Vol. 25, Issue 2, April, 2011, 341-354.

Bazjanac V. (2012). Ifc BIM-Based Methodology for Semi-Automated Building Energy Performance Simulation, Joint conference on information technology for construction and information and knowledge management in building, October 26-28, Sophia Antipolis, France.

Bergés M., Goldman E., Matthews S.H., and Soibelman L. (2008). Training Load Monitoring Algorithms on Highly Sub-Metered Home Electricity Consumption Data," *Tsinghua Science \& Technology*, vol. 13, p. 406, 2008.

Bogen A. C., East W. E., and Rashid M. (2011). A framework for building intelligence fusion, Joint conference on information technology for construction and information and knowledge management in building, October 26-28, Sophia Antipolis, France.

Boll S.F. (1979). Suppression of Acoustic Noise in Speech using Spectral Subtraction, *Journal of IEEE Acoustics Speech and Signal Processing*, Vol. 27, 113-120.

Cook D.J. (2007). Making sense of sensor data, *IEEE Journal of Pervasive Computing*, Vol. 6, 105-108.

Cooley J.W. and Tukey J.W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series, *Journal of Mathematics of Computation*, Vol. 19, 297-301.

Cortes C. and Vapnik V.N. (1995). Support-Vector Networks, *Journal of Machine Learning*, Vol. 20, 273-297.

Day W.H.E. and Edelsbrunner H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods, *Journal of Classification*, Vol. 1, 7-24.

East E.W. (2012-a.) Building Programming information exchange (BPie) ,[Online], Available: http://www.buildingsmartalliance.org/index.php/projects/activeprojects/31 [Accessed: 04 Oct. 2012]

East E.W. (2012-b.) Building Automation Modeling information exchange (BAMie), [Online], Available: http://www.buildingsmartalliance.org/index.php/projects/activeprojects/180 [Accessed: 04 Oct. 2012]

East E.W., Bogen A.C., and Rashid, M., "Life-Cycle Building Control,"9th European Conference on Product and Process Modeling," July 25-27, Reykjavik, Iceland, 2012.

East E., Love D., and Nisbet, N. A Life-Cycle Model for Contracted Information Exchange International Council for Research and Innovation in Building and Construction, Joint Conference on Information Technology for Construction and Information and Knowledge Management in Building, November 16-19 ,Cairo, Egypt, 2010.

Kim H., Stumpf A., and Kim W. (2011) "Analysis of an Energy Efficient Building Design through Data Mining Approach", Automation in Construction, Vol. 20. Issue 1. pages 37-43, Jan. 2011

Kohavi R. And John G.H. (1997). Wrappers for feature subset selection, *Journal of Artificial Intelligence*, Vol. 97, 273-324.

Liu M., CLaridge D.E., Turner W.D. (2002). Continuous Commissioning Guidebook – Maximizing Building Energy Efficiency and Comfort, U.S. Department of Energy, http://www1.eere.energy.gov/femp/pdfs/ccg02_introductory.pdf [Accessed 3 Oct. 2012]

Livingood W., Stein J., Considine T. and Sloup C. Review of Current Data Exchange Practices: Providing Descriptive Data to Assist with Building Operations Decisions, National Renewable Energy Laboratory Technical Report, NREL/TP-5500-50073, May 2011.

MacQueen J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Maile T., Bazjanac V., and Fischer M. (2012). A Method to Compare Simulated and Measured Data to Assess Building Energy Performance. Building and Environment, 56:241–251, October 2012.

McCulloch W.S. and Pitts W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics*, Vol. 5, 115-137.

National Institute of Building Sciences. (2012). Unified Facilities Criteria Program, Whole Building Design Guide – A Program of the National Institute of Building Sciences [Online]. Available: http://www.wbdg.org/references/pa_dod.php [Accessed: 2 Mar. 2012].

Park M.W., Koch C. and Brilakis I. (2012). Three-Dimensional Tracking of Construction Resources using an On-Site Camera System, Journal of Computing in Civil Engineering, American Society of Civil Engineers, Volume 26, Issue 4, July 2012, Pages 541 – 549

Posenato D., Lanata F., Inaudi D., Smith I. (2008). Model-free Interpretation for Continuous Monitoring of Complex Structures, *Advanced Engineering Informatics*, Volume 22, 135-144.

Pradhan A., Ergen E., and Akinci B. (2009). Technological Assessment of Radio Frequency Identification Technology for Indoor Localization, *Journal of Computing in Civil Engineering*, July/August 2009, Vol. 23, Issue 4, 230-238.

Press W.H., Teukolsky S.A., Vetterling W.T., and Flannery B.P. (2007). *Numerical Recipes: The Art of Scientific Computing, 3rd ed*. Cambridge University Press.

Rowe A., Berges M.E., Bhatia G., Goldman E., Rajkumar R., Garrett J.H., Moura J.M.F., and Soibelman L. (2011), Sensor-Andrew: Large-Scale Campus-Wide Sensing and Actuation, *IBM Journal of Research and Development*, Vol. 55, No 2&2, Paper 6.

Saitta S., Raphael B., Smith I.F. (2007). A Bounded Index for Cluster Validity, Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, 174 – 187.

Shahandashti, S.M., Razavi, S.N., Soibelman, L., Berges, M., Caldas, C.H., Brilakis, I., Teizer, J., Vela, P.A., Haas, C., Garrett, J., Akinci, B., and Zhu, Z. (2011). Data Fusion Approaches and Applications for Construction Engineering, Invited paper from article at the 2010 NSF Construction Engineering Conference, Journal of Construction Engineering and Management, American Society of Civil Engineers, Volume 137, Issue 10, October 2011, Pages 863 – 869

Song J., Hass C.T., Caldas C., Ergen E., and Akinci B. (2006). Automating the task of Tracking the Delivery of Receipt of Fabricated Pipe Spools in Industrial Projects, *Automation in Construction*, Volume 15, pages 166-177.

Steinberg A.N., Bowman C.L., and White F.E. (1999). Revisions to the JDL Data Fusion Model, *Proceedings of 1999 SPIE Sensor Fusion: Architectures, Algorithms, and Applications*, vol 3719, 1218-1230.

Tang, P., Huber D., Akinci B., Lipman R., Lytle A. (2010), Automatic Reconstruction of as-built Building Information Models from Laster-Scanned Point Clouds: A Review of Related Techniques, *Automation in Construction*, Vol. 19, no. 7, 829- 843.

Vaseghi S.V. (2009). *Advanced Digital Signal Processing and Noise Reduction*. John Wiley and Sons, Inc.

Yu Z., Fung B.C.M., Haghighat F., Yoshino H., and Morofsky E. (2011). A Systematic Procedure to Study the Influence of Occupant Behavior on Building Energy Consumption, Energy and Buildings, Volume 43, Issue 6, June 2011, Pages1409-1417

Zach R., Glawischnig S., Hönisch M., Appel R., and Mahdavi A. (2012). MOST: An open-source, vendor and technology independent toolkit for building monitoring, data preprocessing, and visualization, 9th European Conference on Product and Process Modeling," July 25-27, Reykjavik, Iceland, 2012.